SENCA-st: Integrating Spatial Transcriptomics and Histopathology with Cross Attention Shared Encoder for Region Identification in Cancer Pathology

Shanaka Liyanaarachchi[†] Chathurya Wijethunga[†] Shihab Aaqil Ahamed[†] Akthas Absar[†]
Ranga Rodrigo[†]
[†]University of Moratuwa, Sri Lanka

Abstract

Spatial transcriptomics is an emerging field that enables the identification of functional regions based on the spatial distribution of gene expression. Integrating this functional information present in transcriptomic data with structural data from histopathology images is an active research area with applications in identifying tumor substructures associated with cancer drug resistance. Current histopathology-spatial-transcriptomic region segmentation methods suffer due to either making spatial transcriptomics prominent by using histopathology features just to assist processing spatial transcriptomics data or using vanilla contrastive learning that make histopathology images prominent due to only promoting common features losing functional information. In both extremes, the model gets either lost in the noise of spatial transcriptomics or overly smoothed, losing essential information. Thus, we propose our novel architecture SENCA-st (Shared Encoder with Neighborhood Cross Attention) that preserves the features of both modalities. More importantly, it emphasizes regions that are structurally similar in histopathology but functionally different on spatial transcriptomics using cross-attention. We demonstrate the superior performance of our model that surpasses state-of-the-art methods in detecting tumor heterogeneity and tumor micro-environment regions, a clinically crucial aspect.

1. Introduction

Understanding structural and functional regions in tissues provides insights of underlying pathophysiology of disease states. While histopathology image analysis helps in understanding structural regions, it lacks representation of functional (spatial) regions. Spatial transcriptomics, on the other hand, enables gene expression visualization in space [4], offering a key advantage over bulk transcriptomics by providing spatially resolved functional information. Identification of functional regions, obscure directly in histopathology images, makes spatial transcriptomics a valuable tool specifically for region segmentation in tumor heterogeneity and tumor micro-environment [13].

The concept of tumor heterogeneity arises when rapidly mutating and dividing cancer cells undergo natural selection [25]. This consists of a tumor edge that conceals the tumor from immune attacks and drug therapies and a tumor core with invasive cancer. Cellular receptors and genetic makeup of the tumor edge are much like the non-cancer cells and make the tumor indistinguishable while the invasive core rapidly undergoes cancer metastasis [34]. The tumor micro-environment refers to the surrounding regions of a tumor, which influence the immune system's ability and the effectiveness of therapeutic drugs in targeting the tumor [15]. The segmentation of these regions, the goal of our work, is crucial in clinical planning and testing experimental therapies.

This region identification is challenging due to the presence of thousands of gene channels and undetermined nature of certain gene expression. Early statistical and machine learning based region segmentation attempts relied solely on spatial transcriptomics (unimodal) [3, 22, 30]. These unimodal architectures turned out to be suboptimal due to transcriptomics inherently being extremely noisy [33]. Leveraging structural features of histopathology images substantially reduce the effects of noisy nature. This combined with the increased importance of both structural and functional regions led to successful multi-modal approaches [33].

Early multi-modal architectures operate at two extremes functional heavy, or structural heavy biasing toward a single modality without a balanced contribution from both modalities. On the functional heavy extreme, SpaGCN [9] uses structural morphologies as weights of the graphs while DeepST [33] uses structural morphologies to augment spatial transcriptomic data. Both involve minimal inclusion of structural information, and models get confused with noisy spatial transcriptomics data resulting in lower performance. On the other structural-heavy extreme, ConGcR/ConGaR [16] uses contrastive learning between spatial transcriptomic embeddings and histopathology image

patch embeddings which cause structural features to dominate the outcome due to the uncontrolled information flow. This drawback calls for a balanced contribution from the special transcriptomic and histopathology image modalities for region segmentation.

In this paper, we present a shared encoder that would learn a fair joint representation of both spatial transcriptomics and histopathology processed in two different branches but control the information flow using a neighborhood cross attention mechanism. This mechanism emphasizes the nodes that have a different correlation ratio of the features of spatial transcriptomics and histopathology compared to their neighbors. This learns a joint representation not dominated by structural features at the local level. We also introduce a hierarchical learning mechanism in which pooled low-resolution data gets trained using contrastive learning, smoothing noise at the global feature level to reduce inherent noise of spatial transcriptomic data but not directly affecting the critical local features. We demonstrate that our model surpasses the state-of-the-art region segmentation results both qualitatively and quantitatively using publicly available data [1, 10, 32]. Our code is available at https://github.com/shanaka-liyanaarachchi/SENCA-st

Our major contributions are

1. Using a novel neighborhood cross-attention shared encoder between histopathology and spatial transcriptomics data leading for better segmentation.
2. Hierarchical learning in which feature flow from histopathology to spatial transcriptomics controlled at according to resolution.

2. Related Works

Machine learning models related to spatial transcriptomics can be divided into two main categories: generative models [6, 17, 31] that predict spatial transcriptomics and inference models that derive insights from spatial transcriptomics data. While many models belong to the first category, our SENCA-st model belongs to the second category of inference models.

Earlier machine learning models existed only using spatial transcriptomics. Using graph neural networks to process spatial transcriptomics data has been suggested in these papers and these papers usually excels at tasks that are strongly presented by multiple gene channels such as segmenting layers of brain cortex. In currently existing multimodal architectures, a single modality is predominantly determining the output features. SpaGCN [9] use structural morphologies as weights of the spatial transcriptomics graphs and DeepST [33] use structural morphologies to augment spatial transcriptomic data. ConGcR [16] changes the approach by using the contrastive learning between two modalities.

Previous work has used graph neural networks maximizing mutual information between local node representations and a global summary and they use contrastive learning between augmented versions of feature vector [3, 38]. As these mechanisms do not effectively mitigate the inherent noisiness of spatial transcriptomics data, Lin *et al*. [16] have suggested using contrastive learning between spatial transcriptomics and histopathology images. However, this approach ultimately results in excessive smoothing, leading to the loss of valuable functional information at the local level. We address this issue by introducing a hierarchical learning mechanism, where smoothing occurs at the global level, while local-level features are learned using a cross-attention shared encoder.

Self-attention was first introduced by Vaswani *et al*. [26], primarily for natural language applications, generating attention weights for language tokens using queries, keys, and values derived from the same language sequence. There is an extended version for graphs [27]. Self attention has been previously used in uni-modal spatial transcriptomics-only models [22, 30]. The concept of cross-attention involves generating queries from one modality while generating keys and values from another modality. Previously, attention has been used to relate between different levels of resolution of histopathology images in generative models [31, 37]. To the best of our knowledge, we are the first to suggest neighborhood cross-attention to learn multi-modal representation between spatial transcriptomics and histopathology for multimodal inference.

3. Method

In our SENCA-st architecture (Fig. 1) we have two separate spatial transcriptomics branch and a histopathology branch to generate RNA embeddings(R) and histopathology image embeddings(H). A shared representation learning(S) of these two branches are learn through the cross attention shared encoder that would fairly represent both modalities but also weigh more attention when structurally similar but functionally different regions. This shared embeddings(S) are used to segment out the regions

First, we create a graph G using spatial locations of the spots in which the gene expression has been measured. Expressed genes are then added as the feature vector V_i of each spot (node) i [3]. Then, we process the graph through a graph transformer g to generate RNA embeddings ($R = g(G)$). Second, in the histopathology image branch, we crop the histopathology image around the spots and pass each image patch through an image encoder to generate image embeddings H. These embeddings run through the cross-attention module. This operation preserves the features of both modalities and emphasizes regions that are structurally similar in histopathology but functionally different in spatial transcriptomics through cross-attention.

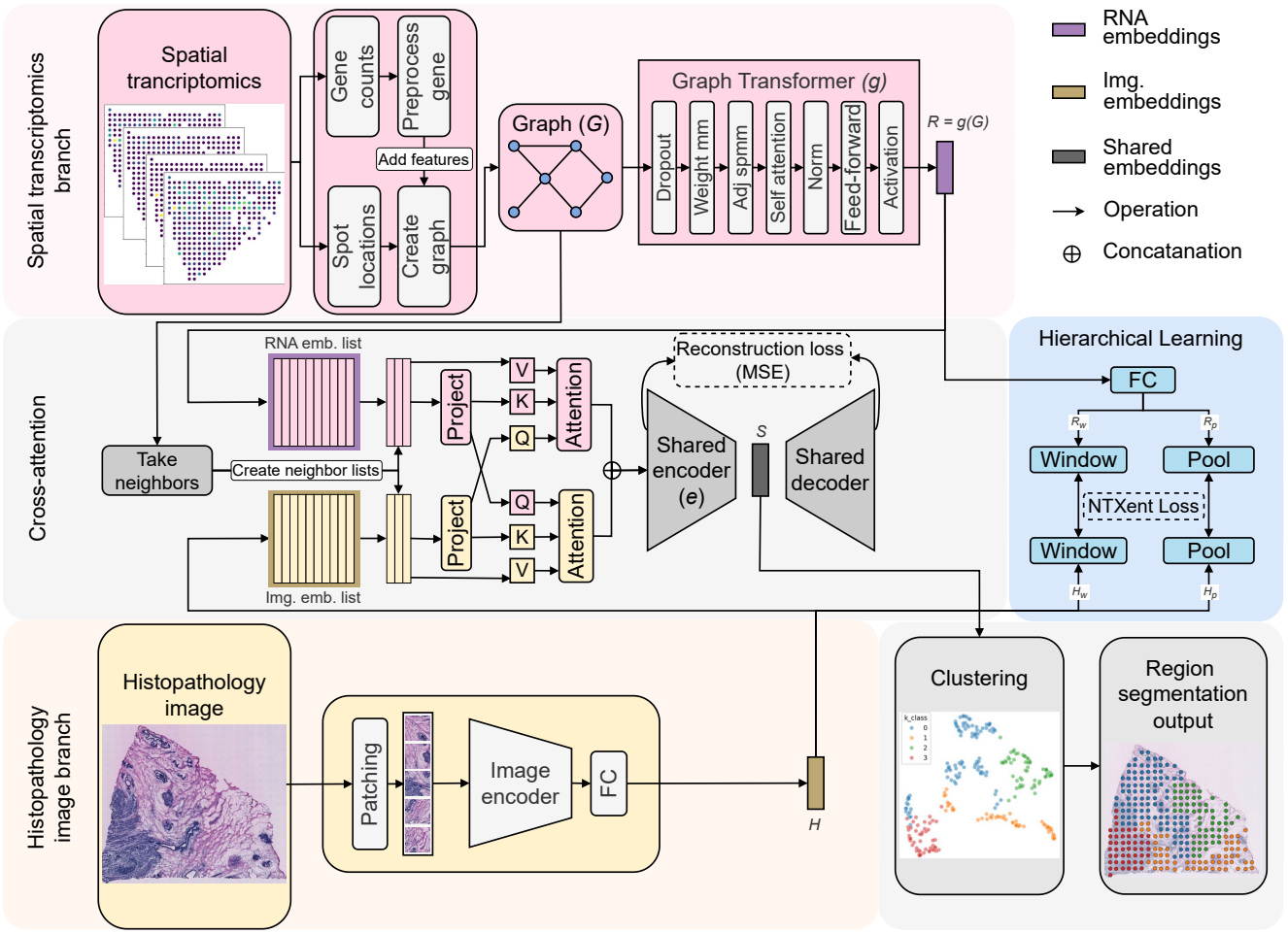

Figure 1. SENCA-st system architecture: we leverage both spatial transcriptomics and histopathology data processed through a graph transformer and a ResNet encoder to produce embeddings through neighborhood cross-attention shared encoder which leads to better segmentation.

Then, we extract the latent space shared embedding S_i. This shared embeddings are clustered and the cluster label of the corresponding spot is used to segment the spots. We use a weighted combination of Mean Square Error (MSE) between encoder input E_i and decoder output D_i with two NT-Xent contrastive losses proposed in SimCLR [5], each for windowed correspondence (H_w, R_w) and pooled correspondence (H_p, R_p) of RNA and image embeddings with λ to adjust weight.

$$\text{Loss} = \text{NT-Xent}_g(H_p, R_p) + \text{NT-Xent}_w(H_w, R_w) + \lambda \underset{i}{\text{avg}}(\text{MSE}(E_i, D_i)). \quad (1)$$

The MSE loss intends to learn local features for spots, while temperature-normalized cross entropy NT-Xent focuses on learning at the global scale. This entire architecture operates in a fully self-supervised and zero-shot fashion. Finally, we benchmark the results using the Adjusted Rand Index (ARI) and compare them to the ground truth label folowing standard becnchmark procedures of literature.

3.1. Graph Creation for Pre-Processing Transcriptomics:

As the first step, we preprocess the spatial transcriptomics data and model it as a graph structure. Following [3], we create the spatial graph G for the spots by considering the nearest neighbors based on their physical distance, using the ball-tree algorithm. Then we preprocess the genes by filtering out those with a total count less than 10, normalizing the gene expression, and transforming it to a log scale. Next, we select the top highly variable genes based on their normalized dispersion similar to ConGCR [16]. We consider these genes as features, and we add the feature vector V_i of each spot i to the graph.

3.2. Transcriptomic Branch—Graph Transformer:

Following the pre-processing step in the spatial transcriptomic branch, we use a graph transformer g to process the graph that models the spatial transcriptomic data to learn a lower-dimensional representation of nodes. The graph transformer begins with a dropout layer, which serves more than just preventing overfitting; it also acts as a masking mechanism for a noisy auto-encoder. Next, we perform a matrix multiplication of learnable weights, followed by another spatial matrix multiplication that also considers the adjacency matrix. This process is similar to the message-passing operation in a graph, where information is exchanged between adjacent nodes learning local features. Then the graph passes the output through a transformer block with self-attention, using the same spot embeddings as queries, keys, and values. This allows the model to focus on important nodes in the same way that the original language transformer pays attention to relationships between word embeddings.

Then, we pass the self attention output through layer normalization and a Feed-Forward (FF) network which plays a more significant role in our model than in the original language model. By having a FF network that runs on every node, it also assigns a weight to each gene, in addition to the weights given to nodes by the self-attention layer. Then, a residual connection bypasses the transformer block similar to the original language model [26]. Next, we pass the embeddings through an activation layer to produce the final output. The graph transformer ensures that it learns a lower-dimensional representation of nodes while considering their spatial positions, and the attention mechanism helps assign more weight to important nodes.

3.3. Histopathology Branch - Image Encoder:

We split the histopathology image into patches by considering a given patch radius (up to third degree neighbor - three times the distance between two spots) around the physical locations of each spot and convert to tensors. Afterwards, we pass these tensors through a ResNet [8] image encoder followed by a Fully Connected (FC) layer to generate patch embeddings. We load the ResNet with pre-trained ImageNet [7] weights, to provide a basic understanding of visual features, although not specific to histopathology. These encoded path embeddings carry information about the structural features of the tissue spot to the shared encoder e.

3.4. Cross-Attention Shared Encoder:

We pass the RNA embedding (R_i) and the image embedding (H_i) of each spot, along with the RNA embeddings (R_n) and image embeddings (H_n) of their neighbors (according to the adjacency matrix) through separate projection layers to generate $H_{n,p}$ and $R_{n,p}$. The two attention layers follow this. For the first attention layer, we generate keys, values, and queries using projected RNA embeddings, original RNA embeddings, and projected image embeddings, respectively. Similarly, for the second attention layer, we generate keys, values, and queries using projected image embeddings, original image embeddings, and projected RNA embeddings, respectively. Then, we generate the cross-attention outputs, which weighs embeddings of spots compared to their neighbors. This results in cross-attention between the transcriptomic and histopathology image modalities. Finally, we extract and concatenate the two vectors for the considered spot. This concatenated vector goes through a dimensionally reducing encoder and then through a dimensionally growing decoder. The encoder output corresponding to each spot, S_i leads to the final segmentation.

$$S_i = e(A_i^1 + A_i^2) \tag{2}$$

$$A_i^1 = \text{softmax}\left((H_{n,p}W^{q1})(R_{n,p}W^{k1})^T/\sqrt{E}\right) R_n W^{v1} \tag{3}$$

$$A_i^2 = \text{softmax}\left((R_{n,p}W^{q2})(H_{n,p}W^{k2})^T/\sqrt{E}\right) H_n W^{v2} \tag{4}$$

($W^{q\cdot}$, $W^{k\cdot}$,$W^{v\cdot}$: weights of queries, keys, values respectively and E: dimension of the keys vector.) We use the shared decoder output to train the model while extracting the latent embedding from the encoder as the shared embedding for clustering.

3.5. Hierarchical Learning:

We leverage hierarchical learning to effectively learn both local (high resolution) and global (low resolution) features. We use NT-Xent-loss-based contrastive learning at a higher level (low resolution) through windowing and pooling, while the cross-attention shared encoder learns local features (high resolution). The shared encoder slides through every spot and learns a local shared embedding, which we use in clustering. Running contrastive learning at a higher level helps smooth out the noisy nature of spatial transcriptomics controllably. However, it does not directly affect the locally learned shared embedding, as it is done at the global scale, unlike in vanilla contrastive models [16]. Finally, we cluster the extracted shared embeddings unsupervised using agglomerative clustering. This approach, which does not use any labels, is important in looking for unknown regions.

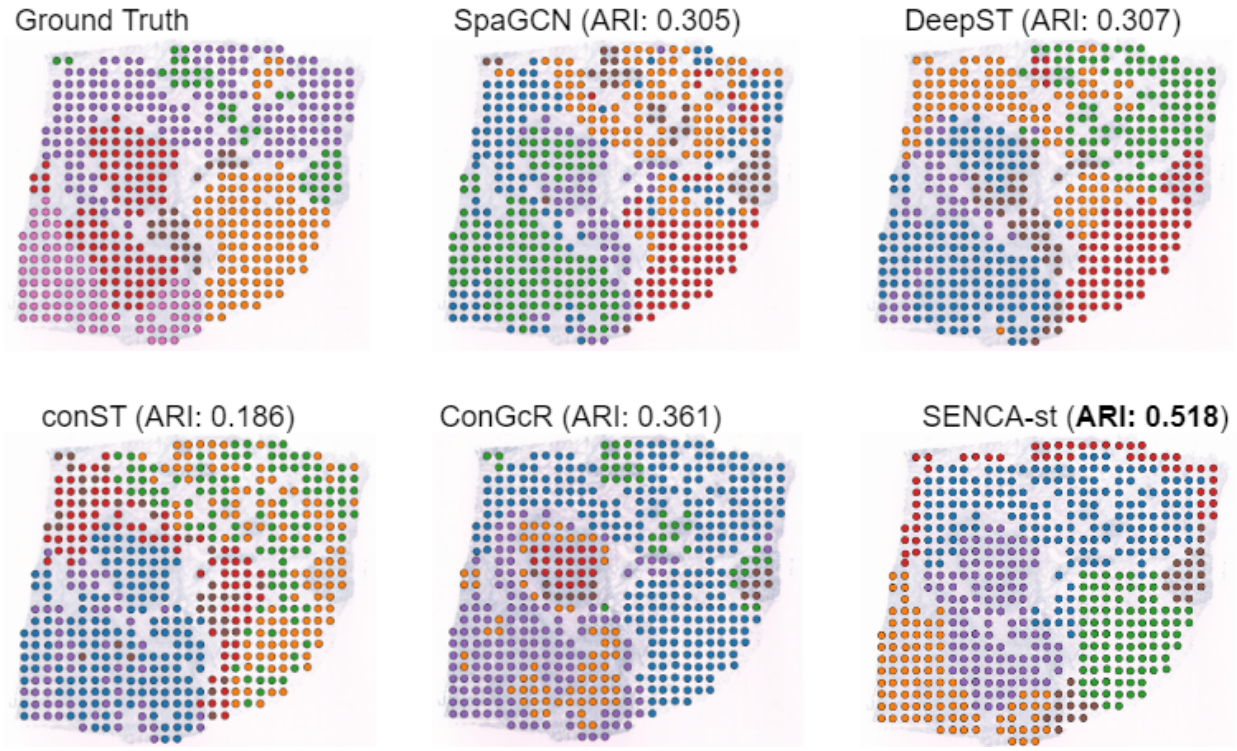

Figure 2. Qualitative comparison for H1 sample in HER2ST: SENCA-st (our) clustering closely matches with the ground truth. For this particular sample ARI is high at 0.518.

Table 1. HER2ST Full Dataset Results: We report the mean and median ARI results (over the eight annotated samples). Our results significantly surpasses the existing results.

Model	ARI (Mean)	ARI (Median)
BayesSpace [35]	0.100	0.071
SpaGCN [9]	0.195	0.230
DeepST [33]	0.237	0.257
conST [38]	0.149	0.111
ConGcR [16]	0.268	0.258
ConGaR [16]	0.187	0.184
SENCA-st (Ours)	**0.304**	**0.320**

4. Experiments and Results

4.1. Region Identification in Breast Cancer

4.1.1. Standardized Testing with HER2ST Dataset

We used publicly available HER2ST dataset[1] from a study by Andersson *et al*. [1] which comprises sections of Human Epidermal Growth Factor Receptor (HER2) positive breast cancer patients. HER2 is a suitable cancer type for targeted therapy [18, 29]) and susceptible to show resistance to targeted therapies [28], where tumor heterogeneity is an important parameter. The dataset contains data for 36 breast cancer sections of 8 patients comprising HE-stained images and spatial transcriptomic for each section. There are 8 samples, one per patient, with pathologist's annotations and scientific explanations, which we use for our experiments following Lin *et al*. [16].

We evaluated all 8 annotated samples independently and calculated the arithmetic mean of the sample ARI (ranges from -1 no agreement to 1 perfect agreement) for the dataset following prior work. ARI measures the similarity between our self-supervised results and ground-truth labels.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\Big/\binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\Big/\binom{n}{2}\right]} \tag{5}$$

For each sample, we train the model for 10 epochs with a learning rate of 5×10^{-4} using the Adam optimizer **without prior training on any other sample**. As model parameters, we used 500 as the number of highly variable genes, 128 as the embedding dimension for RNA and image embeddings, 256 as the hidden dimensions of the graph transformer, 4 as the number of neighbors,$\lambda = 40$ and 0.2 as the dropout rate.

[1]Dataset Downloadable from the official github repository of the study.

Table 2. We conducted several extended experiments on other breast cancer types that were not covered in HER2ST dataset.

Sample Name	(Type)	ARI
CID44971	(TNBC)	0.326
CID4535	(ER+)	0.272
1160920F	(TNBC)	0.300

We train our model on a P100 GPU with 16GB of memory.

According to the benchmark results of the previous models from Lin *et al*. [16], **our model outperforms other methods with a substantial margin** (Table 1). The performance increase in our model is statistically significant ($p < 0.05$) with a p value of 0.0055 for the t-test.

Although the quantitative results offer a fundamental understanding of our model's superior performance, its true value lies in the qualitative analysis of real scenarios. In Fig. 2, the pink cluster in ground truth represents an invasive cancer at the core and the red cluster represents cancer in situ exposing edge. **Ours is the only model that could correctly identify those two regions that are critical in tumor heterogeneity**.

4.1.2. Extended Experiments.

We conducted extended experiments for other breast cancer types as well since the standard dataset only contained HER2+ breast cancer samples. Those samples [2] with more than 7 regions were taken from a study by Wu *et al*. [32]. Those samples used a more information rich technology compared to previous experiment thus had to change embedding dimensions from 256,128 to 512,256 and simultaneously increased input highly variable genes from 500 to 1000. For those samples also each sample, we train the model for 40 epochs with a learning rate of 5×10^{-4} using the Adam optimizer **without prior training on any other sample**. We report the results in table 2.

4.2. Region Identification in Squamous Cell Carcinoma

We devised our SENCA-st model to identify spatial regions of Squamous Cell Carcinoma samples[3] from a study by Ji et. al. [10]. For this experiment also we set model parameters similar to the previous experiment and trained the model for 40 epochs with a learning rate of 5×10^{-4} using the Adam optimizer **without prior training on any other sample**. We used statistical testing using the Wilcoxon test to find genes correlated with the clusters, and the clusters were extremely correlated with known marker genes.

[2]Dataset Downloadable from the official zenodo repository of the study.
[3]Dataset Downloadable from the official NCBI GEO of the study.

Table 3. Marker genes of patient 2 sample of Squamous Cell Carcinoma detected with Wilcoxon test performed to test the gene is strongly correlated to the particular cluster compared to the rest

Gene Name	Cluster	P-value
KRT2	0	4.771447×10^{-11}
FLG	0	2.019676×10^{-07}
KLK7	0	1.831156×10^{-03}
DSC1	0	8.774093×10^{-03}
MMP9	1	1.095943×10^{-10}
MMP1	1	1.095943×10^{-10}
CCL21	1	2.575223×10^{-08}
MMP3	1	1.880689×10^{-07}
DCD	2	5.048839×10^{-38}
IGFBP5	2	5.310009×10^{-31}
DCN	2	1.315863×10^{-18}
KRT2	3	5.668670×10^{-14}
COL1A1	3	1.488934×10^{-08}
COMP	3	1.488934×10^{-08}
LOR	3	1.542446×10^{-07}
SPRR1B	4	2.752813×10^{-35}
S100A7	4	1.199169×10^{-31}
SPRR1A	4	1.281949×10^{-30}
SBSN	4	1.844224×10^{-29}
DSC2	4	4.377951×10^{-27}
IGFBP4	5	4.068289×10^{-18}
CCDC80	5	8.295423×10^{-17}
DIO2	5	8.428684×10^{-11}
FAAP20	6	6.318033×10^{-05}
CASP14	6	3.996734×10^{-03}

This also means **that our model could be used to identify biomarkers of unknown pathologies since we did not use any supervision or previous knowledge on the fact that those gene channels are biomarkers when training the model**.

In the experiment with the patient 2 sample correlation with identified marker genes with Wilcoxon test is reported in 'Table 3 and visualization of identified marker genes along with identified regions are presented figure 3. Cluster 0 (blue) is statistically significantly($p < 0.05$) correlated with KRT2, KRTDAP that are responsible genes for differentiation of keratinocytes [2]. Cluster 0 also had cancer biomarkers such as KLK7 [12], DSC1 [20] statistically significantly correlated to it (Table 3). Cluster 4 (purple) had cancer biomarkers such as SPRR1B [19], SPRR1A (suggesting tumor is under immune attack with CAF (Cancer-Associated Fibroblast) [14](Table 3). Overall both cluster 0 and 4 had SBSN [36] cancer biomarker (Fig 3) but statistically more significantly in cluster 4 (Table 3). Overall cluster 0 and 4 are two distinctive genetically heterogeneous

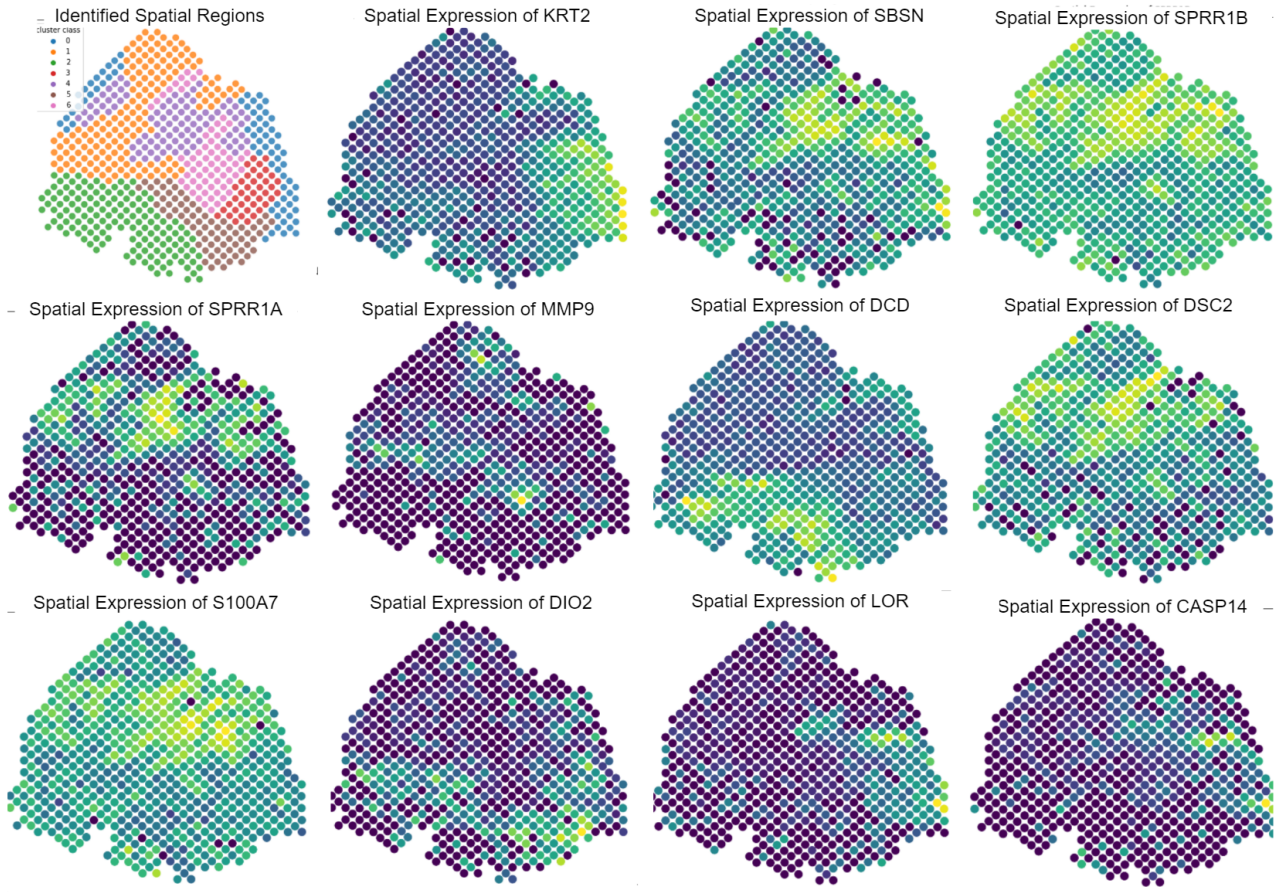

Figure 3. Spatial Clusters of sample p2 of Squamous Cell Carcinoma with marker genes.

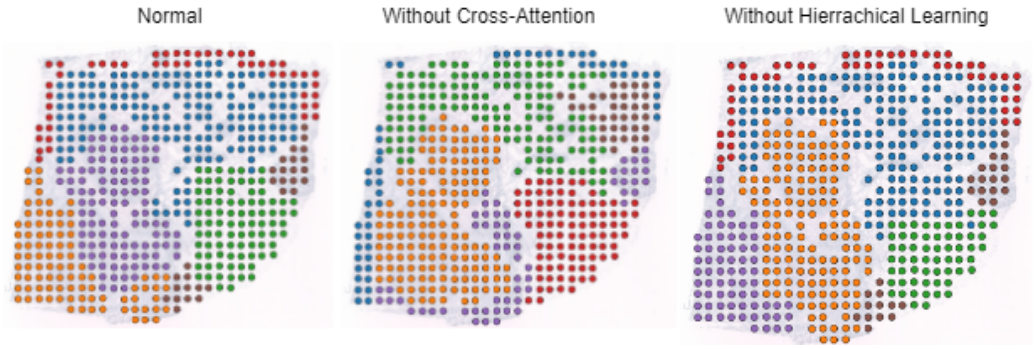

Figure 4. Ablation Study - Qualitative effect visualization on H1 sample of HER2ST dataset. We investigated effect of components of the system by removing them and conducting ablation studies.

cancer regions.

Cluster 1 (orange) had MMP9, MMP3, MMP1 which are of matrix metalloproteinase (MMP) family that are involved immune activation in tumor micro-environment [11] especially in immune cell recruitment such as Tumor-Associated Macrophages. Cluster 2 (green) has a strong correlation (Table 3) with DCD [23] which a gene prominently expressed in sweat glands and also some variants suggests cachexia (cancer related muscle degradation) [24]. Cluster 3(red) has a strong correlation with LOR as well as KRT2 and Cluster 5(brown) has strong correlation with DIO2 [21].

This experiment signifies that our model is capable in

Table 4. Ablation Studies - Quantitative Results. We conducted ablation experiments isolating key components of the system by removing them and benchmarking with the HER2ST dataset to study their effect.

Ablation Experiment	ARI (Mean)
Normal	0.3039
Without Cross-Attention Weights	0.2218
Without Hierarchical Learning	0.2495

Table 5. Effect of number of neighbors. Even though a small change from actual closest neighbors does not deviate results, large changes make the model confused. This indicate the importance of neighbors

n	n=4	n=8	n=16
ARI(Mean)	0.304	0.290	0.240

understanding extremely complicated underlying disease pathology phenomena without getting caught to the inherent noise of spatial transcriptomics. We have covered the full scope of experimentation with achieving the state of the art results with a strong margin in test cases where standard performance indices with annotations are available and statistically showcasing the ability for a self supervised model to understand structural and functional regions statistically matching the existing literature.

4.3. Ablation Studies

We conducted ablation studies in-order to isolate contribution of each part of the proposed system. Without the cross attention weights ARI droped to 0.2218 and without Hierarchical Learning ARI dropped to 0.2495 from the original 0.3039 signifying the importance of these components as reported in table 4. Qualitatively without cross attention model was not able to identify structurally similar functionally different regions effectively, when the hierarchical learning was removed cross attention identified the regions but their borders were not close to ground truth like in the original with both the components as seen in the figure 4.

4.4. Model Parameters

We studied the effect of number of neighbors since the spatial transcriptomic grid pattern is uniform square as in Fig 2 number of closest neighbors is 4 which is the recommended number. However when it is increased to 8 including diagonal neighbors other than the orthogonal neighbors results was not deviated much. But when number was increased to 16 which goes beyond immediate neighborhood, results started deviating. We report the results in Table 5

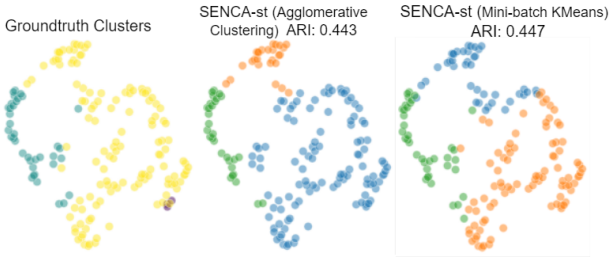

Figure 5. Clustering of shared embeddings generated by SENCA-st of C1 sample of HER2ST Dataset

4.5. Clustering

There are several limitations in clustering that could be addressed in future work. First being Optimal number of clusters, even-though we benchmark for annotations for currently known regions there might be regions yet to be discovered and in the Figure 5 we could see some more clusters than the ground truth of three in the UMAP(Uniform Manifold Approximation and Projection). Other limitation is being class size bias in which the small purple cluster not being identified by the algorithm even though embeddings of these spots have been aggregated together. We keep the agglomerative clustering as it produces a dendogram which could be used identify more substructures.

5. Conclusion

In this paper, we propose SENCA-st, a neighborhood cross-attention-based shared encoder architecture integrating spatial transcriptomics and histopathology image data. Cross-attention assigns weights to structurally similar but functionally different regions and demonstrates how hierarchical learning diffuses structural features at low resolution without affecting important local functional regions. These novelties have contributed to our architecture performing better than existing work quantitatively and qualitatively.

We test our model quantitatively with standard benchmarking achieving SOTA performances as well as statistically validating the correlation between identified regions with known biomarkers. More importantly our model demonstrated capabilities to identify critical regions that were not possible to detect with previous methods. Overall we present a novel architecture that address critical issues in existing literature and achieving SOTA results both quantitatively and qualitatively. Future work could focus on the clustering limitations discussed above. We hope that our contributions will have a significant impact on the molecular pathology research community.

Acknowledgments - This project was partially supported by Accelerating Higher Education Expansion and Development (AHEAD) Operation funded by the World Bank.

References

[1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12 (1):6012, 2021. 2, 5

[2] Balvinder K Bloor, Nicholas Tidman, Irene M Leigh, Edward Odell, Bilal Dogan, Uwe Wollina, Lucy Ghali, and Ahmad Waseem. Expression of keratin k2e in cutaneous and oral lesions: association with keratinocyte activation, proliferation, and keratinization. *The American journal of pathology*, 162(3):963–975, 2003. 6

[3] Zixuan Cang, Xinyi Ning, Annika Nie, Min Xu, and Jing Zhang. SCAN-IT: Domain segmentation of spatial transcriptomics images by graph neural network. In *In Proceedings of the British Machine Vision Conference*, page 406, 2021. 1, 2, 3

[4] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 3

[6] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[9] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021. 1, 2, 5

[10] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *cell*, 182(2): 497–514, 2020. 2, 6

[11] Kai Kessenbrock, Vicki Plaks, and Zena Werb. Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell*, 141(1):52–67, 2010. 7

[12] Simon Kind, Carolina Palacios Castillo, Ria Schlichter, Natalia Gorbokon, Maximilian Lennartz, Lisa S Hornsteiner, Sebastian Dwertmann Rico, Viktor Reiswich, Florian Viehweger, Martina Kluth, et al. Klk7 expression in human tumors: a tissue microarray study on 13,447 tumors. *BMC cancer*, 24(1):794, 2024. 6

[13] Sabrina M Lewis, Marie-Liesse Asselin-Labat, Quan Nguyen, Jean Berthelet, Xiao Tan, Verena C Wimmer, Delphine Merino, Kelly L Rogers, and Shalin H Naik. Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9):997–1012, 2021. 1

[14] Jing Li, Ling-Long Tang, and Jun Ma. Survival-related indicators alox12b and sprr1a are associated with dna damage repair and tumor microenvironment status in hpv 16-negative head and neck squamous cell carcinoma patients. *BMC cancer*, 22(1):714, 2022. 6

[15] Qiongyu Li, Xinya Zhang, and Rongqin Ke. Spatial transcriptomics for tumor heterogeneity analysis. *Frontiers in Genetics*, 13:906158, 2022. 1

[16] Yu Lin, Yanchun Liang, Duolin Wang, Yuzhou Chang, Qin Ma, Yan Wang, Fei He, and Dong Xu. A contrastive learning approach to integrate spatial transcriptomics and histological images. *Computational and Structural Biotechnology Journal*, 23:1786–1795, 2024. 1, 2, 3, 4, 5, 6

[17] Gabriel Mejia, Daniela Ruiz, Paula Cárdenas, Leonardo Manrique, Daniela Vega, and Pablo Arbeláez. Enhancing gene expression prediction from histology images with spatial transcriptomics completion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 91–101. Springer, 2024. 2

[18] Diogo Mendes, Carlos Alves, Noémia Afonso, Fátima Cardoso, José Luís Passos-Coelho, Luís Costa, Sofia Andrade, and Francisco Batel-Marques. The benefit of HER2-targeted therapies on overall survival of patients with metastatic HER2-positive breast cancer–a systematic review. *Breast Cancer Research*, 17:1–14, 2015. 5

[19] Yoshitaka Michifuri, Yoshihiko Hirohashi, Toshihiko Torigoe, Akihiro Miyazaki, Jyunki Fujino, Yasuaki Tamura, Tomohide Tsukahara, Takayuki Kanaseki, Junichi Kobayashi, Takanori Sasaki, et al. Small proline-rich protein-1b is overexpressed in human oral squamous cell cancer stem-like cells and is related to their growth through activation of map kinase signal. *Biochemical and Biophysical Research Communications*, 439(1):96–102, 2013. 6

[20] MP Myklebust, Ø Fluge, H Immervoll, A Skarstein, L Balteskard, O Bruland, and O Dahl. Expression of dsg1 and dsc1 are prognostic markers in anal carcinoma patients. *British journal of cancer*, 106(4):756–762, 2012. 6

[21] A Nappi, C Miro, AG Cicatiello, S Sagliocchi, L Acampora, F Restolfer, and M Dentice. The thyroid hormone activating enzyme, dio2, is a potential pan-cancer biomarker and immunotherapy target. *Journal of Endocrinological Investigation*, 48(5):1149–1172, 2025. 7

[22] Till Richter, Anna Schaar, Francesca Drummer, Cheng-Wei Liao, Leopold Endres, and Fabian J Theis. SpatialSSL: Whole-brain spatial transcriptomics in the mouse brain with self-supervised learning. In *NeurIPS 2023 AI for Science Workshop*, 2023. 1, 2

[23] Birgit Schittek, Rainer Hipfel, Birgit Sauer, Jürgen Bauer, Hubert Kalbacher, Stefan Stevanovic, Markus Schirle,

Kristina Schroeder, Nikolaus Blin, Friedegund Meier, et al. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. *Nature immunology*, 2(12):1133–1137, 2001. 7

[24] Grant D Stewart, Richard JE Skipworth, James A Ross, Kenneth CH Fearon, and Vickie E Baracos. The dermcidin gene in cancer: role in cachexia, carcinogenesis and tumour cell survival. *Current Opinion in Clinical Nutrition & Metabolic Care*, 11(3):208–213, 2008. 7

[25] Tuomas Tammela and Julien Sage. Investigating tumor heterogeneity in mouse models. *Annual Review of Cancer Biology*, 4(1):99–119, 2020. 1

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[27] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. 2

[28] Claudio Vernieri, Monica Milano, Marta Brambilla, Alessia Mennitto, Claudia Maggi, Maria Silvia Cona, Michele Prisciandaro, Chiara Fabbroni, Luigi Celio, Gabriella Mariani, et al. Resistance mechanisms to anti-HER2 therapies in HER2-positive breast cancer: Current knowledge, new research directions and therapeutic perspectives. *Critical Reviews in Oncology/Hematology*, 139:53–66, 2019. 5

[29] Jiani Wang and Binghe Xu. Targeted therapeutic options and future perspectives for HER2-positive breast cancer. *Signal transduction and targeted therapy*, 4(1):34, 2019. 5

[30] Tianqi Wang, Huitong Zhu, Yunlan Zhou, Weihong Ding, Weichao Ding, Liangxiu Han, and Xueqin Zhang. Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. *Communications Biology*, 7(1):1351, 2024. 1, 2

[31] Xiaofei Wang, Xingxu Huang, Stephen Price, and Chao Li. Cross-modal diffusion modelling for super-resolved spatial transcriptomics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 98–108. Springer, 2024. 2

[32] Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021. 2, 6

[33] Chang Xu, Xiyun Jin, Songren Wei, Pingping Wang, Meng Luo, Zhaochun Xu, Wenyi Yang, Yideng Cai, Lixing Xiao, Xiaoyu Lin, et al. DeepST: Identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research*, 50(22):e131–e131, 2022. 1, 2, 5

[34] Qichao Yu, Miaomiao Jiang, and Liang Wu. Spatial transcriptomics technology in cancer research. *Frontiers in Oncology*, 12:1019111, 2022. 1

[35] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology*, 39(11):1375–1384, 2021. 5

[36] Zhongqiu Zhou, Zhuojun Zhang, Han Chen, Wenhao Bao, Xiangqin Kuang, Ping Zhou, Zhiqing Gao, Difeng Li, Xiaoyi Xie, Chunxiao Yang, et al. Sbsn drives bladder cancer metastasis via egfr/src/stat3 signalling. *British Journal of Cancer*, 127(2):211–222, 2022. 6

[37] Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, et al. Asign: an anatomy-aware spatial imputation graphic network for 3d spatial transcriptomics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30829–30838, 2025. 2

[38] Yongshuo Zong, Tingyang Yu, Xuesong Wang, Yixuan Wang, Zhihang Hu, and Yu Li. conST: An interpretable multi-modal contrastive learning framework for spatial transcriptomics. *BioRxiv*, pages 2022–01, 2022. 2, 5